Tagging Enriched Bank Transactions Using LLM-Generated Topic Taxonomies

Daniel de S. Moraes
TeleMídia Lab - PUC-Rio
danielmoraes@telemidia.puc-rio.br

Polyana B. da Costa
TeleMídia Lab - PUC-Rio
polyana@telemidia.puc-rio.br

Pedro T. C. Santos
TeleMídia Lab - PUC-Rio
thiagocutrim@telemidia.puc-rio.br

Ivan de J. P. Pinto
TeleMídia Lab - PUC-Rio
ivan@telemidia.puc-rio.br

Sérgio Colcher
TeleMídia Lab - PUC-Rio
colcher@inf.puc-rio.br

Antonio J. G. Busson
BTG Pactual
antonio.busson@btgpactual.com

Matheus A. S. Pinto
BTG Pactual
matheus.adler@btgpactual.com

Rafael H. Rocha
BTG Pactual
rafael-h.rocha@btgpactual.com

Rennan Gaio
BTG Pactual
rennan.gaio@btgpactual.com

Gabriela Tourinho
BTG Pactual
gabriela.tourinho@btgpactual.com

Marcos Rabaioli
BTG Pactual
marcos.rabaioli@btgpactual.com

David Favaro
BTG Pactual
david.favaro@btgpactual.com

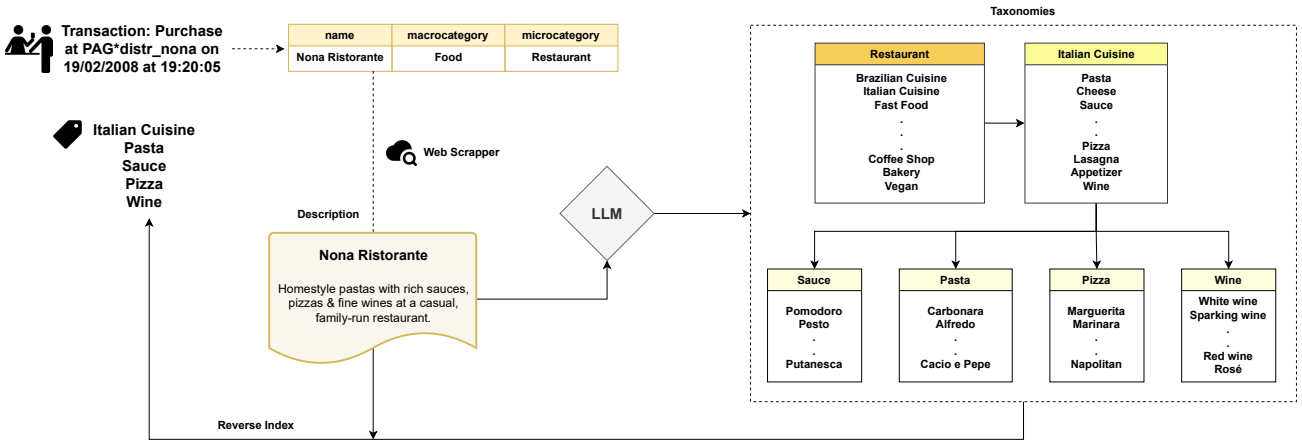

Figure 1: Overview of our method. (Note: Merchant and transaction data are fabricated for demonstration only).

ABSTRACT

This work presents an unsupervised method for tagging banking consumers' transactions using automatically constructed and expanded topic taxonomies. Initially, we enrich the bank transactions via web scraping to collect relevant descriptions, which are then preprocessed using NLP techniques to generate candidate terms. Topic taxonomies are created using instruction-based fine-tuned LLMs (Large Language Models). To expand existing taxonomies with new terms, we use zero-shot prompting to determine where to add new nodes. The resulting taxonomies are used to assign descriptive tags that characterize the transactions in the retail bank dataset. For evaluation, 12 volunteers completed a two-part form assessing the quality of the taxonomies and the tags assigned to merchants. The evaluation revealed a coherence rate exceeding 90% for the chosen taxonomies. Additionally, taxonomy expansion using LLMs demonstrated promising results for parent node prediction, with F1-scores of 89% and 70% for *Food* and *Shopping* taxonomies, respectively.

KEYWORDS

Large Language Models, Natural Language Processing, Web Scrapping, Topic Modeling

1 INTRODUCTION

Many recent studies have focused on the application of Machine Learning-based methods for the classification and characterization of financial transactions. For example, Vollset et al. [23] and Busson et al. [4] explored an approach to hierarchically classifying

In: Proceedings of the Brazilian Symposium on Multimedia and the Web (WebMedia'2024). Juiz de Fora, Brazil. Porto Alegre: Brazilian Computer Society, 2024.

© 2024 SBC – Brazilian Computing Society.

ISSN 2966-2753

financial transactions, employing a predefined set of categories/subcategories that describe purchase types and transactions. However, these methods apply a limited, predefined set of static classes, which restricts the ability to extend classifications based on user experiences when encountering new, previously undefined categories.

In this context, to expand the possible set of classes/tags to label a transaction, we developed an unsupervised method based on *topic taxonomies*. Taxonomies are very useful in the structural and semantic analyses of topics and textual data. However, creating and maintaining them is often costly and challenging to scale manually. Therefore, recent works have tackled the automatic creation and expansion of *topic taxonomies*, in which each node in a hierarchy represents a conceptual topic composed of semantically coherent terms.

We present an unsupervised method for automatically constructing and expanding topic taxonomies with instruction-based LLMs (Large Language Models), in a Zero-Shot manner. Candidate terms for the initial version of the taxonomy are obtained using topic modeling and keyword extraction techniques. Then we apply LLMs to post-process the resulting terms, create a hierarchy, and add new terms to an existing taxonomy. Since the taxonomies are derived from a corpus of unstructured texts describing niches of consuming habits, we opted to investigate the use of LLMs in our approach. LLMs are often pre-trained on a large corpus of text, allowing them to learn contextual representations that capture the intricacies of human language.

We applied our method to a private dataset of transactions of a retail bank, enriched with scraped data from food and shopping companies, and evaluated the resulting taxonomies quantitatively. The generated tags of our topic taxonomies are then assigned to the bank transactions characterizing the companies in each transaction, as shown on Figure 1. In total, 58 topic taxonomies were created for the *Food* category and 6 for the *Shopping* category.

A two-step quantitative evaluation was conducted on a subset of the taxonomies. For this evaluation, we selected the topic taxonomies with the highest number of terms in each category: "Brazilian Cuisine" from *Food* and "Clothing and Accessories" from *Shopping*. Taxonomies with more terms are most likely to result in a deeper hierarchy, which gives more data for evaluation. We asked 12 volunteers to answer a two-part form, which assessed the quality of the created taxonomies and the quality of the tags assigned to label transactions. The evaluation showed an average coherence of tags to transactions above 90%.

As more scraped data from food and shopping companies are added to the retail bank's dataset, the topic taxonomies will need to be updated to include new terms. We used LLMs for this task as well, employing commercial LLMs like Gemini Pro [1] and GPT-4 [14], alongside open-source LLM options such as LLaMA-Alpaca (7B) [22], Phi-2[1], and Mixtral 8x7B [9]. We showcase their results in both taxonomy creation and expansion. For the expansion part, we also compared our method to existing ones (a BERT-based method and Musubu[20]) on the SemEval dataset and our generated taxonomies as well. Gemini Pro achieved the best results, with F1-scores of 89% and 70% for parent node prediction on the *Food* and *Shopping* taxonomies, respectively.

[1]https://huggingface.co/microsoft/phi-2

The remainder of the paper is structured as follows: Section 2, reviews the related work, highlighting existing approaches. Section 3 provides the necessary background, laying the foundation for our methodologies and contextualizing our contributions. Section 4 details the dataset construction process, explaining how we enriched and prepared the data for the taxonomies' construction. In Section 5, we describe the creation of the taxonomies, outlining the methods used to generate them. Section 6 discusses the expansion of the taxonomies, demonstrating how they can be dynamically extended to accommodate new categories. Section 7 focuses on the evaluation of these taxonomies, presenting the metrics and results that validate their accuracy and also the quality of the tags assigned to the transactions. Finally, Section 8 concludes the paper, summarizing our findings, discussing their implications, and suggesting directions for future research.

2 RELATED WORK

Taxonomies represent the structure behind a collection of documents, organizing the hierarchical relationships between terms in a tree structure [13]. They play an essential part in the structural and semantic analysis of textual data, providing valuable content for many applications that involve information retrieval and filtering, such as web searching, recommendation systems, classification, and question answering.

Since creating and maintaining taxonomies is a costly task, often difficult to scale if done manually, methods that automatically construct and update them are desirable. Early works on automatic taxonomy creation focused on building hypernym-hyponym taxonomies, where each pair of terms expresses an 'is-a' relationship [19]. More recent works have tackled the automatic creation of other taxonomies, such as topic taxonomies. In a topic taxonomy, each node represents a conceptual topic composed of semantically coherent terms.

In this context, Zhang et al. [28] developed TaxoGen, an unsupervised method for constructing topic taxonomies. Taxogen uses the SkipGram model from an input text corpus to embed all the concept terms into a latent space that captures their semantics. In this space, the authors applied a clustering method to construct a hierarchy recursively based on a variation of the spherical K-means algorithm.

Another work that focuses on topic taxonomies is TaxoCom [10], a framework for automatic taxonomy expansion. TaxoCom is a hierarchical topic discovery framework that recursively expands an initial taxonomy by discovering new sub-topics. It uses locally discriminative embeddings and adaptive clustering, resulting in a low-dimensional embedding space that effectively encodes the textual similarity between terms. One main disadvantage of TaxoCom is that it requires a large set of quality phrases in the target language, and curating these phrases can be costly. The quality of the output taxonomy is highly dependent on those phrases.

Regarding the automatic expansion of taxonomies, an important related example is Musubu [20], a framework for low-resource taxonomy enrichment that uses a Language Model (LM) as a knowledge base to infer term relationships. For the taxonomy expansion part of out method, we used Musubu as a baseline for comparison.

As to using Large Language Models for taxonomy tasks, Chen et al. [6] investigated how LLMs, like GPT-3, perform in taxonomy construction tasks. The authors compared two approaches: fine-tuning, which involves training the LLM on a specific dataset to adapt it for taxonomy tasks, and prompt techniques, where the LLM receives instructions and examples to perform a task without being explicitly trained for it. Their findings showed that prompt techniques such as few-shot learning generally outperform fine-tuning, particularly with smaller datasets. Based on these findings, we applied prompting techniques, specifically zero-shot prompting, across various LLMs to assess their effectiveness in constructing and expanding taxonomies. Section 7 shows the results of our approach, as well as the results of applying Musubu[20] as baseline.

3 BACKGROUND

In this section, we provide a comprehensive background on Large Language Models (LLMs), and the concept of Prompt-tuning. These concepts are essential to understanding the construction and editing of taxonomies utilizing LLMs.

3.1 Large Language Models

Lately, Large Language Models (LLMs) have garnered significant attention for their exceptional performance in various NLP tasks. LLMs, such as GPT-3[3] and LLAMA[22], are characterized by their massive scale, comprising billions of parameters and being trained on vast amounts of data. These models are often pre-trained in an unsupervised manner on large corpora of textual data, such as books, articles, and web pages, allowing them to learn contextual representations that capture the intricacies of human language.

To use LLMs for specific purposes, a highly effective approach is to fine-tune them on task-specific data. Fine-tuning enables LLMs to adapt to specific domains or tasks with minimal labeled data, significantly reducing the need for large annotated datasets. However, in scenarios where labeled data is scarce or difficult to obtain, LLMs can also be used without specific training or additional data, in a Zero-Shot manner [21]. Given the scale of these models and the data they are trained on, LLMs embed vast knowledge that enables them to achieve high generalization capabilities and perform tasks in diverse contexts, even without specific training for those tasks[16].

In our experiments, we tested several types of language models, from private LLMs (GPT 4 [14], Gemini Pro[2]), to open-source LLMs (Llama 2 [22]), to a Mixture of Experts LLM (Mixtral [9], and a Small Language Model (SLM), Phi 2 [3].

3.2 Prompt Engineering

Prompt Engineering is a fundamental technique used to enhance the performance and adaptability of Large Language Models (LLMs) in specific tasks or domains [7]. It involves optimizing and crafting prompts to efficiently use language models (LMs) [3]. This approach allows researchers and practitioners to tailor the behavior and output of LLMs, making them more suitable for targeted applications.

Techniques such as Zero-shot prompting, Few-shot prompting, Chain of Thought, ReAct, Self-Consistency etc. have been explored to guide LLMs toward desired responses [18, 21, 25–27]. The effectiveness of prompt tuning has been demonstrated in various applications, including question-answering, summarization, and dialogue generation. The choice of prompt greatly influences the generated output, and by carefully crafting prompts, researchers can guide the model's responses toward desired behaviors. For example, in language translation, a prompt can specify the source language and desired target language to ensure accurate and fluent translations. In our method, we used the Zero-Shot prompting technique.

3.3 Zero-Shot Prompting

Since LLMs (Large Language Models) are trained on vast amounts of data, they can follow instructions and perform tasks in contexts where they were not specifically trained, in a Zero-Shot (ZS) manner. This prompting style allows the model to adapt, making it versatile. A Zero-Shot (ZS) prompt directly instructs the model to perform a task without additional examples or demonstrations to guide the LLM's response, which is why they are also known as task instructions [21].

In a study by Li [11], the authors highlighted several advantages of using ZS prompts, such as the ability to craft highly interpretable prompts, requiring fewer training data or examples, a more straightforward prompt design process, and a flexible prompt structure. Additionally, Reynolds and McDonell [17] noted that carefully engineered zero-shot prompts can outperform few-shot prompts in certain scenarios, as examples can sometimes be interpreted as part of a narrative rather than as a guiding mechanism. This finding also influenced our decision to use zero-shot prompting in our method.

4 DATASET CONSTRUCTION

This work uses a proprietary dataset consisting of consumer transactions from a retail bank. Each transaction includes only a merchant name indicating the business where that purchase occurred along with macro and micro categories as illustrated in Figure 1 The macro and micro are originally assigned by [4] using the information from the business activities and products.

We focus on two macro-categories from this dataset: *Food* and *Shopping*, selecting the top 50,000 businesses with the highest number of transactions for each category.

With the limited initial information, assigning detailed tags to transactions is challenging. To address this, we augment the dataset through a data enrichment process involving web scraping. Using tools such as Selenium[4] and Beautiful Soup[5], we gathered activity descriptions for companies in each macro category. For the *Food* macro category, the search was conducted on specialized platforms for restaurants and food delivery services. For the *Shopping* macro category, we obtained establishment descriptions directly from internet indexing and search tools.

In the context of enrichment for the *Food* macro category, web scraping was conducted as follows: (1) the centers of all Brazilian state capitals and the Federal District were used as base locations

[2]https://deepmind.google/technologies/gemini/pro/

[3]https://www.microsoft.com/en-us/research/blog/phi-2-the-surprising-power-of-small-language-models/

[4]https://www.selenium.dev/about/

[5]https://readthedocs.org/projects/beautiful-soup-4/downloads/pdf/latest/

for restaurant searching; (2) for each location, restaurants listed on the first one hundred pages of the platform were extracted. After completing these steps, the information was combined with the merchant database using the merchant's name and micro categories.

For the *Shopping* macro category, the description of each merchant was obtained using Google Search Engine[6], selecting descriptions from the first ten search results. The final description consists of a concatenation of all the obtained descriptions. The search queries were constructed using the merchant names combined with their micro categories.

5 TAXONOMY CONSTRUCTION

To automatically create topic taxonomies for *Food* and *Shopping* businesses, we developed a 3-step method. First, we preprocess the descriptions in our enriched dataset to retain only the relevant parts of the text. Next, we apply two techniques to select candidate terms for the topic taxonomies: keyword extraction and topic modeling.

In the post-processing phase, we use large language models (LLMs) to refine the results of each step, filtering out unrelated terms. Finally, we use LLMs again to organize the final terms into hierarchies, forming the topic taxonomies.

5.1 Preprocessing

We applied a few NLP techniques to refine the businesses' descriptions in our dataset. At first, we remove stop words to eliminate commonly used words that do not carry significant meaning in our contexts. Then, to retain only the most relevant portions of the descriptions, we employ part-of-speech (POS) tagging to identify and exclude words that belong to specific POS categories. The list of POS tag categories that were removed includes ADV, CCONJ, ADP, AUX, CONJ, DET, INTJ, PART, PRON, PUNCT, SYM, SCONJ, ADJ, VERB, PROPN.[7]

After this initial preprocessing step, we run the first iteration of the candidate term selection part to build a filter of generic words, not to create topic taxonomies yet. For this step, we use the entire corpus of descriptions for each macro category, resulting in two corpora (*Food* and *Shopping*). For each micro category in the macro categories' corpora, we use Keyword Extraction and Topic Modeling to gather candidate terms for the filter, combining the results of both techniques in a list. Then, We use an LLM to remove the terms it identifies as unrelated to the main topic (each micro category) from the list. The prompt that we used for requesting this separation is illustrated below.

```
prompt= "Given the terms in the following list: "+
     <wordsList> +". Separate them into two groups. In
     group 1 the terms with no relation to the topic "+
     <type> +". And in group 2 the terms that are related."
```

Listing 1: Prompt for separating candidate terms related to the type of establishment

By using this prompt, we try to ensure that the model's response is consistently formatted according to the pattern described in it, facilitating the processing of the resulting string, although, some of the LLMs we tested did not output the response in the requested format. Once we complete one iteration of this method for each macro category in our dataset, we add the words of group 2 to the corresponding list of generic words. We apply the corresponding filter of generic words for each macro category corpus, resulting in the final preprocessed corpus.

5.2 Candidate Terms Selection

For this part of our method, we use each preprocessed corpus separately. For the *Food* corpus, we group the descriptions based on their micro-categories, creating 58 sub-corpus specific to that domain. We have six micro categories for the *Shopping* corpus, resulting in 6 specific sub-corpus. The candidate terms selection methods are applied to each sub-corpus, creating topic taxonomies where the main topic is the micro category.

5.2.1 Keyword Extraction. The first approach to candidate term selection was to use an unsupervised keyword selection method called Yake! [5]. This method is based on statistical text features extracted from single documents to select the most relevant keywords from that text. It does not require training on a document set and is not dependent on dictionaries, text size, language, domain, or external corpora.

Yake! allows for the specification of parameters such as the language of the text, the maximum size of the n-grams being sought, and others. In our method, we customized only the language to Portuguese, and the maximum number of keywords sought for each set of descriptions was 30 words.

After extracting the keywords from each group of descriptions, we obtained a total set of N candidate terms. However, these terms are further filtered using an LLM, where we ask it to separate the terms related to the main topic from those unrelated, as explained earlier in subsection 5.1.

5.2.2 Topic Modeling. Our second approach to collecting initial topics and candidate terms was Topic Modeling. We applied the Latent Dirichlet Allocation algorithm [2], available at the Gensim Library[8].

We construct a dictionary for each macro-category corpus in our macro-categories corpora by extracting unique tokens and bigrams. After a few empirical tests, we set the minimum frequency of a bigram to 20 occurrences. Since some corpora have a minimal number of tokens (the micro category "Greek Cuisine" from the *Food* macro category has only five stores marked as such, with a corpus of only 127 tokens), we had to set a reasonably small number so that smaller corpora could also have a few bigrams. With the resulting dictionary of tokens, the LDA algorithm was applied. Three main parameters are to be defined in an LDA algorithm: number of topics, *alpha*, and *beta*.

The number of topics defines the latent topics to be extracted from the corpus. The parameter *alpha* is *a priori* belief in document-topic distribution, while *beta* is *a priori* belief in topic-word distribution.

To define the number of topics for each micro category corpus, we tried numbers from 1 to 5, constantly checking which configuration would result in the best average topic coherence for that

[6]https://www.google.com

[7]https://spacy.io/usage/linguistic-features#pos-tagging

[8]https://pypi.org/project/gensim/

corpus. Small corpora would have 1 or 2 topics, while bigger ones would have 5. To correctly define the *alpha* and *beta* priors, we would have to analyze the distribution for each category corpus [24]. Since this would be rather difficult, we set those priors to be auto-defined by the LDA algorithm, which learns these parameters based on the corpus. We select the terms with the highest coherence with the resulting topics. Each topic returns 20 words with their coherence scores, but we do not use all of them as some have very low coherence. After testing a few configurations, for each topic, we select 60% of the terms with the highest coherence within that topic.

With initial terms for each topic taxonomy, we ask an LLM to separate the ones closely related to the main topic from those unrelated, as mentioned earlier.

5.3 Hierarchy Construction

Once we have the post-processed lists of candidate terms obtained by each technique mentioned in subsection 5.2, we merge them and remove repetitions. After the merge, for each macro category, we have lists of terms for each micro category, representing each topic taxonomy. However, they do not have any hierarchy level between the terms configuring the taxonomy.

To tackle this problem, we use an LLM again, this time with a prompt that searches for sub-categories within the terms of a topic to create these hierarchies. The prompt is illustrated below:

```
prompt="Create a dictionary by hierarchically arranging the
    following words:" + <wordsList> +." Use JSON format as
    the output such as the following: {\"key\": [\" list
    of words\"]}"
```

Listing 2: Prompt for creating a hierarchy for each list of tags.

With this prompt, we seek to ensure that the LLM response has a consistent pattern and facilitates handling the returned string. After this step, we have a hierarchy of terms in each topic taxonomy in the *Food* and *Shopping* macro categories.

5.4 Merchant Tagging

With the topic taxonomies for both *Food* and *Shopping* macro-categories, we can now assign tags to merchants/establishments. To do so, we use the descriptions attached to these establishments, and we see which terms from a taxonomy are mentioned in their descriptions with a reverse index algorithm. We employ the taxonomy whose topic is the same as the establishment's micro category, as shown in Figure 2.

6 TAXONOMY EXPANSION

Another essential part of our method is the automatic expansion of existing taxonomies as new terms arrive, derived from additional merchant scrapped data, as shown in Section 4. In this section, we present our approach to taxonomy expansion by using instruction-based LLMs.

As new transactions may include new businesses, new terms can emerge from the descriptions obtained through the scraping process. Therefore, we need to update the taxonomies with these new terms maintaining and enriching the created hierarchies with the potential new terms.

After completing the transaction enrichment process, including the search for business descriptions and the selection of candidate terms, if relevant terms not included in the current hierarchies are detected, we initiate the expansion process.

6.1 Prompt engineering instruction for taxonomy representation

First, we represent our topic taxonomies in a format that can be interpreted by an LLM. We employed a generic prompt, illustrated below, across all tested methods to convert topics into root nodes and their terms into child nodes.

```
Childs of [ROOT]: [CHILD1,CHILD2,CHILD3]
Childs of [CHILD1]: [CHILD4,CHILD5]
Childs of [CHILD2]: [CHILD6]
...
```

Listing 3: Prompt for representation of taxonomy

6.2 Predicting the parent of a node

To experiment with taxonomies expansion, we used two datasets: our *Food* and *Shopping* topic taxonomies and the taxonomies from SemEval-2015 Task 17 [15]. Those are low-resource taxonomies, with thousands of nodes or less, which are appropriate for the current prompt size of LLMs. We used the SemEval dataset to compare the results with well-established methods for taxonomy expansion, such as Musubu [20]. Similar to their experiments, we hid 20% of the terms (chosen randomically) in the taxonomies to predict their respective parent nodes. To verify the parent/root of a new term, we used the following prompt:

Listing 4: Prompt for searching for a node's parent

```
prompt="Who is the father of "+<new_term>+"?"
```

In Table 1, we see the F1-Scores for parent node prediction. Equation 1 showcases how to calculate the F1-Score. TP is the number of true positives, nodes that were correctly assigned as parents of child nodes. FP is the number of false positives, nodes that were incorrectly assigned as a parent to a child node. FN is the number of false negatives, nodes that should have been assigned as parent nodes but were not.

$$F1 = \frac{2 * TP}{2 * TP + FP + FN} \tag{1}$$

For baseline models, we used Bert and Musubu; for commercial LLMs, Gemini Pro and GPT-4; and for open-source LLMs, LLama-Alpaca(7B), Phi-2, and Mixtral 8x7B. We evaluate them in 4 taxonomies from the SemEval dataset and our taxonomies. For each taxonomy, the LLMs perform significantly better than Musubu, with GPT-4 and Gemini Pro having the highest F1-Scores, with the latter beating the former by a few points. However, the most recent open-source options (Phi-2 and Mixtral 8x7B) are getting close in performance.

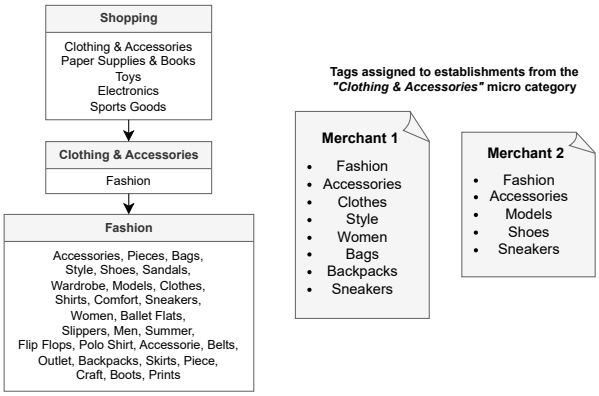

Figure 2: Assigning tags to establishments based on a topic taxonomy.

It is important to note that while SemEval taxonomies have thousands of nodes, ours have only a few hundred, which we can assume is a significant reason for the degrading performance of Musubu and Bert (LMs or LM-based methods). In contrast, the LLMs have a robust performance in such low-resource settings. This also shows that LLMs have a remarkable understanding of questions and zero-shot performance, generalizing well even for datasets in different languages.

7 TAXONOMY EVALUATION

To properly evaluate the topic taxonomies that we created in this work, we developed a two-step qualitative evaluation of a limited part of the results.

In total, 58 topic taxonomies were created for the *Food* set and 6 for the *Shopping* set. For our evaluation, we selected the topic taxonomies with the highest number of terms in each part (the "Brazilian Cuisine" taxonomy for the *Food* part and the "Clothing and Accessories" taxonomy for the *Shopping* one). First, we assess the quality of removing generic terms from each taxonomy, and then, we evaluate the tags assigned to establishments based on that taxonomy. We asked 12 volunteers to answer a two-part form.

Part 1 - Accuracy of the terms that were selected as related to the topic: In this part, we evaluate if the LLMs could correctly group the relevant and non-relevant terms, removing the generic terms. To do so, we defined a ground truth with the relevant terms as true positives and the non-relevant terms as true negatives. Table 3 shows the results.

GPT-4 was the best model, followed by Gemini Pro, both scoring over 60% accuracy for the Brazilian Cuisine taxonomy and over 86% accuracy for the Clothing and Accessories taxonomy. Smaller language models such as Phi 2 and Llama 2 7B performed poorly both in removing generic terms and in formatting the response accordingly, with Phi 2 being particularly verbose.

Part 2 - Human Evaluation of the Quality of the Tagging Process: In this part, the volunteers were asked to examine if the tags assigned to an establishment were appropriate and coherent to that establishment's description. We selected the top 5 establishments with the highest transactions for each micro category. We asked our evaluators to analyze the tags assigned to describe that establishment and choose the ones that were not appropriate. This way, we have a coherence ratio for each establishment based on the number of proper tags divided by the total number of tags. We average the results of our 12 evaluators and present them in Table 2. Figure 2 shows the "Clothing & Accessories" taxonomy that was evaluated and 2 of the merchants and the tags assigned to them that were included in the evaluation.

8 CONCLUSION

In this work, we presented an unsupervised method for automatically creating and expanding topic taxonomies using LLMs. We evaluated some of the generated taxonomies and applied them in transaction tagging in a retailer's bank dataset. The evaluation showed promising results, with average coherence scores above 90% for the two selected taxonomies. The taxonomies' expansion with Gemini Pro also showed exciting results for parent node prediction, with F1-scores of 89% and 70% for *Food* and *Shopping* taxonomies, respectively.

For future work on taxonomy construction, we plan to test more robust term selection methods, such as embedding-based approaches. We also plan on conducting ablation studies to validate whether the keyword extraction and topic modeling parts help improve the quality of the taxonomies created, by using a baseline prompt to ask the LLM to generate child nodes given a parent node. In terms of taxonomy expansion, there are several tasks to explore, ranging from node-level operations to generating entire sub-trees and identifying similar structures. Additionally, we intend to enhance our instruction-tuned LLM for taxonomy tasks

Method	SemEval-2015 Task 17				Our taxonomies	
	Chemical	Equipment	Food	Science	Food	Shopping
Gemini Pro	**0.68**	**0.80**	**0.91**	**0.72**	**0.89**	**0.73**
GPT-4	0.65	0.78	0.89	0.70	0.87	0.71
Mixtral-8x7B	0.59	0.63	0.80	0.57	0.74	0.60
Phi-2	0.56	0.52	0.68	0.56	0.64	0.54
LLama 7B	0.51	0.42	0.58	0.46	0.60	0.49
Musubu	0.35	0.46	0.37	0.42	0.21	0.13
Bert-Base	0.13	0.14	0.12	0.16	0.11	0.06

Table 1: F1-score for parent node prediction.

	Brazilian Cuisine Taxonomy		Clothing & Accessories Taxonomy	
	Average Coherence	Number of Tags	Average Coherence	Number of Tags
Merchant 1	92.30%	10	97.11%	8
Merchant 2	94.23%	8	83.07%	5
Merchant 3	89.23%	5	94.38%	5
Merchant 4	87.17%	6	93.84%	5
Merchant 5	93.40%	7	97.43%	6

Table 2: Results of evaluating the tags assigned to each merchant/establishment.

	Brazilian Cuisine	Clothing & Accessories
Llama 2 7B	29.54%	52.78%
Phi 2	40.90	73.68%
Mixtral 8x7B v0.1	46.93%	70.27%
Gemini Pro	61.36%	86.11%
GPT 4	**68.08**%	**86.84**%

Table 3: Accuracy of using each LLM to remove generic words from each topic taxonomy.

by fine-tuning or employing more efficient methods such as LoRA [8].

LIMITATIONS

To address the limitations of our work, we begin with the taxonomy construction component. In this phase, we relied on topic modeling and keywords extraction to select candidate terms for our taxonomies. The LDA algorithm used for topic modeling performs suboptimally when the base corpus is small. Some of our topics had corpora with vocabularies of fewer than 100 words, which can result in topics containing irrelevant or incoherent terms. Additionally, we could have further experimented with the LDA hyperparameters for each micro-category corpus.

Regarding the evaluation of the generated taxonomies, we did not assess topic completeness. Without a ground truth, it is challenging to quantify how comprehensively the terms in a taxonomy cover the main topic. Furthermore, we evaluated only 2 of the 64 taxonomies generated by our method, leaving a substantial portion unexamined.

In the taxonomy expansion experiment, we evaluated only a low-resource setting with fewer than a thousand nodes. Most studies focus on taxonomies with hundreds of thousands or more nodes. This presents a challenge for LLMs due to their limited context. Addressing this contextual limitation could benefit from insights found in other works that tackle similar issues [12].

ETHICS STATEMENT

In this work, we ensure the utmost protection of customers and store sensitive data by exclusively using non-sensitive information in our dataset. Our prompts solely rely on selected words from store descriptions, thus avoiding any direct usage of personal or sensitive information. No customer-specific data or store-sensitive details are integrated into the system, upholding privacy and security as top priorities.

Moreover, we strictly adhere to ethical guidelines during our experiments involving volunteers, and no personal data is collected from them. Our focus lies solely on analyzing the results of our proposed approach. Participants' anonymity and confidentiality are maintained throughout the research process, ensuring a responsible and trustworthy approach to data handling.

ACKNOWLEDGMENTS

The authors would like to acknowledge BTG Pactual for the partnership and financial support to this research.

REFERENCES

[1] Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models.

[2] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3, Jan (2003), 993–1022.

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[4] Antonio J. G. Busson, Rafael Rocha, Rennan Gaio, Rafael Miceli, Ivan Pereira, Daniel de S. Moraes, Sérgio Colcher, Alvaro Veiga, Bruno Rizzi, Francisco Evangelista, Leandro Santos, Fellipe Marques, Marcos Rabaioli, Diego Feldberg, Debora Mattos, João Pasqua, and Diogo Dias. 2023. Hierarchical Classification of Financial Transactions Through Context-Fusion of Transformer-based Embeddings and Taxonomy-aware Attention Layer. In *Anais do II Brazilian Workshop on Artificial Intelligence in Finance (BWAIF 2023) (BWAIF 2023)*. Sociedade Brasileira de Computação. https://doi.org/10.5753/bwaif.2023.229322

[5] Ricardo Campos, Vítor Mangaravite, Arian Pasquali, Alípio Jorge, Célia Nunes, and Adam Jatowt. 2020. YAKE! Keyword extraction from single documents using multiple local features. *Information Sciences* 509 (2020), 257–289.

[6] Boqi Chen, Fandi Yi, and Dániel Varró. 2023. Prompting or Fine-tuning? A Comparative Study of Large Language Models for Taxonomy Construction. In *2023 ACM/IEEE International Conference on Model Driven Engineering Languages and Systems Companion (MODELS-C)*. IEEE, 588–596.

[7] Sabit Ekin. 2023. Prompt engineering for ChatGPT: a quick guide to techniques, tips, and best practices. *Authorea Preprints* (2023).

[8] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685* (2021).

[9] Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088* (2024).

[10] Dongha Lee, Jiaming Shen, SeongKu Kang, Susik Yoon, Jiawei Han, and Hwanjo Yu. 2022. TaxoCom: Topic Taxonomy Completion with Hierarchical Discovery of Novel Topic Clusters. In *Proceedings of the ACM Web Conference 2022*. 2819–2829.

[11] Yinheng Li. 2023. A Practical Survey on Zero-Shot Prompt Design for In-Context Learning. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*. 641–647.

[12] Xinnian Liang, Bing Wang, Hui Huang, Shuangzhi Wu, Peihao Wu, Lu Lu, Zejun Ma, and Zhoujun Li. 2023. Unleashing Infinite-Length Input Capacity for Large-scale Language Models with Self-Controlled Memory System. *arXiv preprint arXiv:2304.13343* (2023).

[13] Irina Nikishina, Varvara Logacheva, Alexander Panchenko, and Natalia Loukachevitch. 2020. RUSSE'2020: Findings of the First Taxonomy Enrichment

Task for the Russian language. *arXiv preprint arXiv:2005.11176* (2020).
[14] OpenAI. 2023. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL]
[15] Octavian Popescu and Carlo Strapparava. 2015. Semeval 2015, task 7: Diachronic text evaluation. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*. 870–878.
[16] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research* 21, 1 (2020), 5485–5551.
[17] Laria Reynolds and Kyle McDonell. 2021. Prompt programming for large language models: Beyond the few-shot paradigm. In *Extended abstracts of the 2021 CHI conference on human factors in computing systems*. 1–7.
[18] Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. 2024. A systematic survey of prompt engineering in large language models: Techniques and applications. *arXiv preprint arXiv:2402.07927* (2024).
[19] Rion Snow, Daniel Jurafsky, and Andrew Ng. 2004. Learning syntactic patterns for automatic hypernym discovery. *Advances in neural information processing systems* 17 (2004).
[20] Kunihiro Takeoka, Kosuke Akimoto, and Masafumi Oyamada. 2021. Low-resource taxonomy enrichment with pretrained language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 2747–2758.
[21] Adrian Tam. [n. d.]. What Are Zero-Shot Prompting and Few-Shot Prompting. https://machinelearningmastery.com/what-are-zero-shot-prompting-and-few-shot-prompting/. Accessed: 2024-07-02.
[22] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).
[23] Erlend Vollset, Eirik Folkestad, Marius Rise Gallala, and Jon Atle Gulla. 2017. Making use of external company data to improve the classification of bank transactions. In *Advanced Data Mining and Applications: 13th International Conference, ADMA 2017, Singapore, November 5–6, 2017, Proceedings 13*. Springer, 767–780.
[24] Hanna Wallach, David Mimno, and Andrew McCallum. 2009. Rethinking LDA: Why priors matter. *Advances in neural information processing systems* 22 (2009).
[25] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171* (2022).
[26] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, H Chi, Quoc V Le, and Denny Zhou. [n. d.]. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. ([n. d.]).
[27] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. ReAct: Synergizing Reasoning and Acting in Language Models. In *International Conference on Learning Representations (ICLR)*.
[28] Chao Zhang, Fangbo Tao, Xiusi Chen, Jiaming Shen, Meng Jiang, Brian Sadler, Michelle Vanni, and Jiawei Han. 2018. TaxoGen: Unsupervised Topic Taxonomy Construction by Adaptive Term Embedding and Clustering. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (London, United Kingdom) *(KDD '18)*. Association for Computing Machinery, New York, NY, USA, 2701–2709. https://doi.org/10.1145/3219819.3220064